Published in: De Smedt, Koenraad/Hajič, Jan/Kübler, Sandra (eds.): The Sixth International Workshop on Treebanks and Linguistic Theories (TLT '07). Bergen, Norway. December 7–8, 2007. – Tartu: Northern European Association for Language Technology, 2007. Pp. 115-126. (NEALT Proceedings Series 1)

Why is it so difficult to compare treebanks? TIGER and TüBa-D/Z revisited

Ines Rehbein and Josef van Genabith
Dublin City University
School for Computing

Abstract

This paper is a contribution to the ongoing discussion on treebank annotation schemes and their impact on PCFG parsing results. We provide a thorough comparison of two German treebanks: the TIGER treebank and the TüBa-D/Z. We use simple statistics on sentence length and vocabulary size, and more refined methods such as perplexity and its correlation with PCFG parsing results, as well as a Principal Components Analysis. Finally we present a qualitative evaluation of a set of 100 sentences from the TüBa-D/Z, manually annotated in the TIGER as well as in the TüBa-D/Z annotation scheme, and show that even the existence of a parallel subcorpus does not support a straightforward and easy comparison of both annotation schemes.

1 Introduction

Currently, three treebanks are available for German: NEGRA, TIGER (using a slightly improved version of the NEGRA annotation scheme) and TüBa-D/Z. The annotation schemes of the first two treebanks are quite similar, while both differ considerably from TüBa-D/Z. All three corpora contain text from the same domain (newspaper text, but from different newspapers) and use the same POS tag set (Schiller et al., 1995), but there are crucial differences concerning the linguistic theory underlying the syntactic annotation.

The merits and drawbacks of the different annotation schemes and their impact on PCFG parsing constitute an open research question (Kübler et al., 2006; Rehbein & van Genabith, 2007). While Kübler et al. (2006) argue that the TüBa-D/Z is more adequate for PCFG parsing than the NEGRA annotation scheme, based on around 16% better `PARSEVAL` parsing results for a parser trained on the TüBa-D/Z, Rehbein & van Genabith (2007) present experiments that show that the claim does not hold, as PARSEVAL is highly sensitive to the ratio of non-terminal vs. terminal nodes in the trees. However, the question how to compare different treebank annotation schemes in a fair and unbiased way remains unanswered. There are a number of attempts, based on statistical measures, to compare syntactic structure in different corpora: Nerbonne and Wiersma (2006) present an

aggregate measure of syntactic distance based on POS trigrams, Sanders (2007) uses Leaf-Ancestor path based permutation tests to measure differences between dialectal variations of British English. Corazza et al. (to appear) describe a measure based on conditional cross-entropy to predict parsing performance.

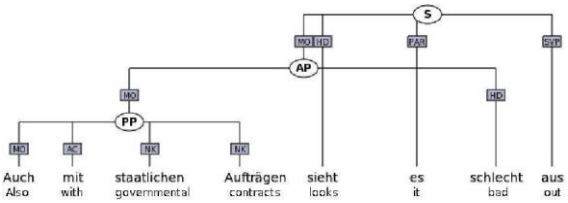
In this paper we take a close look at the similarities and differences between the TIGER and TüBa-D/Z treebanks and annotation schemes, using simple statistics like sentence or word length, vocabulary size as well as more sophisticated methods like Principal Component Analysis and perplexity. We investigate the correlation between perplexity and parsability of a corpus and present a qualitative evaluation of a set of 100 sentences from the TüBa-D/Z, manually annotated in the TIGER as well as in the TüBa-D/Z annotation scheme.

The paper is structured as follows: Section 2 gives an overview of the main features of the two treebanks. Section 3 reports on similarities and differences between the two treebanks. In Section 4 we discuss correlations between corpus homogeneity and parsing results. Section 5 gives a qualitative analysis of the parser output, and the last section concludes.

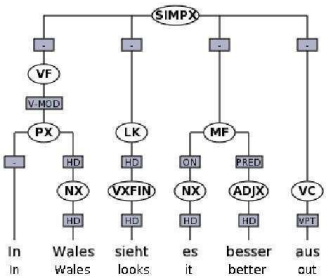
2 The TIGER Treebank and the TüBa-D/Z

The two treebanks used in our experiments are the TIGER treebank (Release 2) (Brants et al., 02) and the Tüba-D/Z (Release 2) (Telljohann et al., 05). TüBa-D/Z consists of approximately 22,000 sentences, while the TIGER Treebank is larger with more than 50,000 sentences. Both treebanks contain German newspaper text and are annotated with phrase structure and dependency (functional) information. Both treebanks use the STTS POS Tag Set (Schiller et al., 95). TIGER uses 44 different grammatical function labels, while TüBa-D/Z utilises only 40 function labels. For the encoding of phrasal node categories TüBa-D/Z uses 26 different categories, TIGER uses a set of 25 category labels. Other major differences between the two treebanks are: in TIGER long distance dependencies are expressed through crossing branches (Figure 1), while in TüBa-D/Z the same phenomenon is expressed with the help of grammatical function labels.

The annotation in the TIGER treebank is rather flat and allows no unary branching, whereas the nodes in TüBa-D/Z do contain unary branches and a more hierarchical structure, resulting in a much deeper tree structure than the trees in the TIGER treebank. This results in an average higher number of nodes per sentence for the TüBa-D/Z (Table 1). Figure 1 illustrates the different annotation of PPs in both annotation schemes: in TIGER the internal structure of the PP is flat and the adjective and noun inside the PP are directly attached to the PP, while TüBa-D/Z is more hierarchical and inserts an additional NP node.



Auch mit staatlichen Aufträgen sieht es schlecht aus.
"It also looks bad for public contracts."



In Wales sieht es besser aus.
"Things seem better in Wales."

Figure 1: TIGER and TüBa-D/Z treebank tree

Another major difference is the annotation of topological fields (Höhle, 1998) in TüBa-D/Z, a descriptive model which captures the semi-free word order in German. Depending on the sentence configuration (verb first, verb second or verb last) the verb can fill in the left (LK) or the right sentence bracket (VC), while the other constituents are ordered relative to the verb in the initial field (VF), the middle field (MF) and the final field (NF).

3 Comparing the Treebanks

We divided both treebanks into sets of samples without replacement with 500 sentences each, randomly selected from the two treebanks, which resulted in 100 samples for the TIGER treebank and 44 samples for the TüBa-D/Z. In order to account for the different size of the treebanks we used samples 1-44 from the TüBa-D/Z treebank as well as samples 1-44 (TIGER1) and 45-88 (TIGER2) from the TIGER treebank.

As we are interested in the influence of sampling techniques on parsing results we also generated a second set of samples with 500 trees each, which were taken in sequential order from the treebanks (rather than randomly as in the first set described above). This means that, in contrast to the random samples, the content in each sample is "semantically" related, which most obviously must have crucial impact on vocabulary size and homogeneity of the samples.

3.1 Sentence Length / Word Length / Vocabulary Size

The average sentence length in TIGER is comparable to the one in TüBa-D/Z (Table 1), but the average word length in TüBa-D/Z is shorter than in TIGER. TüBa-D/Z also uses a smaller vocabulary than the TIGER treebank. Due to the flat annotation in TIGER the ratio of non-terminal vs. terminal nodes is smaller than in TüBa-D/Z. While the treebanks are comparable with regard to text domain and sentence length, there are considerable differences concerning word length and

vocabulary size between the two corpora. In the next section we investigate the distribution of POS tags in TIGER and TüBa-D/Z, using Principal Components Analysis.

	avg. sent. length (rand)	avg. word length (rand)	avg. vocab size (rand)	avg. vocab size (seq)	non-term. /term. nodes
TIGER1	17.86	6.27	2992	2638	0.47
TIGER2	17.03	6.27	2989	2662	0.47
TüBa-D/Z	17.25	5.70	2906	2585	1.20

Table 1: Some properties of the TIGER and TüBa-D/Z treebank

3.2 Principal Component Analysis (PCA) of POS Tags

PCA is a way of reducing complex, high-dimensional data and detecting underlying patterns by transforming a high number of (possibly) correlated variables in a multivariate data set into a smaller number of uncorrelated variables whilst retaining as much as possible of the variation present in the data. The uncorrelated new variables are called principal components or *eigenvectors*. They are chosen in such a way that high correlating variables are combined into a new variable which describes the largest part of the variance in the data. The new variable constitutes the first principal component. Next the second component is chosen so that it describes the largest part of the remaining variance, and so on. PCA has been successfully applied to a number of tasks such as the analysis of register variation (Biber, 1998) or authorship detection (Juola & Baayen, 1998).

Figure 2 shows the 1st and 2nd components of a PCA based on the frequency counts of POS tags in the randomised samples, which together capture around 33% of the variance in the data. The first component clearly separates TIGER from TüBa-D/Z samples. TüBa-D/Z is characterised by a high number of informal elements such as interjections, foreign language material (mostly anglicisms), indefinite and interrogative pronouns and indicators of a personal style such as personal pronouns. TIGER samples show a high number of nouns, determiners, attributive adjectives, prepositions and also circumpositions, past participles and first elements of compounds. A high number of nominal elements (nouns, compounds, nominalised adjectives) is typical for a nominative style (Ziegler et al., 2002), which is often regarded as being more objective and informative than a verbal style. Due to space constraints we can only offer a preliminary analysis: we tend to interpret the first componont as a dimension of informality, where formal texts with a high degree of information content are positioned at one end and informal texts written in a more personal and subjective style at the other end.

3.3 Perplexity

Kilgariff (2001) describes how the information-theoretic measure of *cross-entropy* can be used to assess the *homogeneity* of a text corpus. Perplexity is the log of the

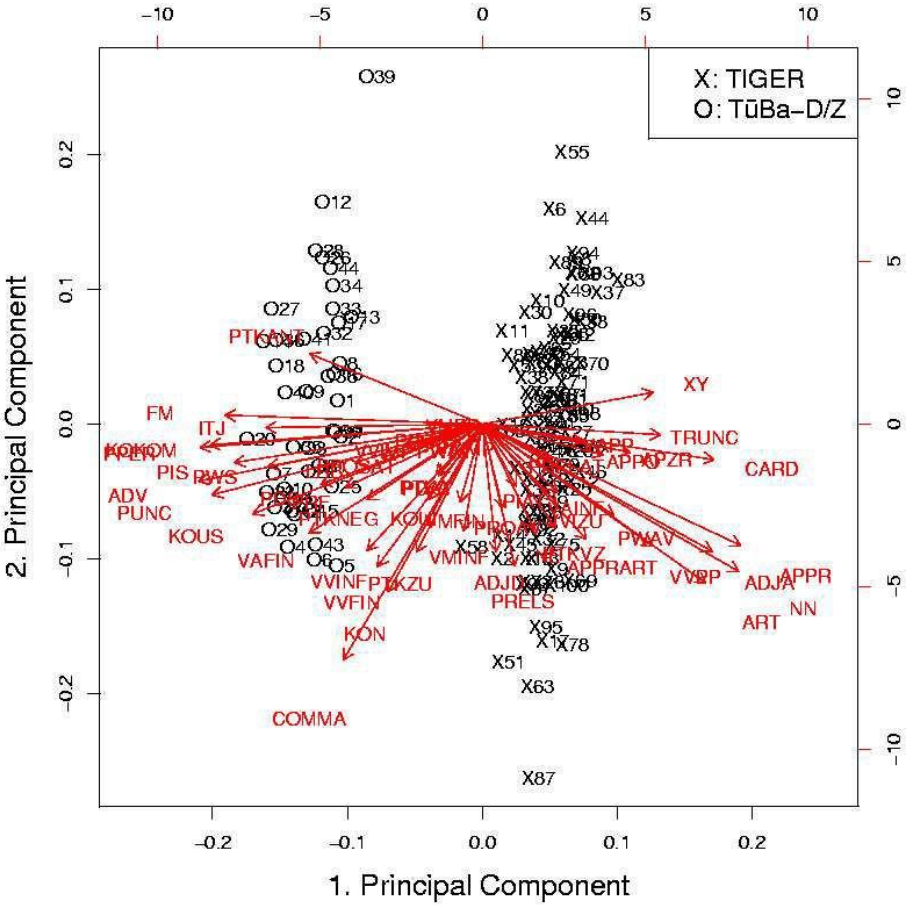


Figure 2: PCA for TIGER/TüBa-D/Z POS tags

cross-entropy of a corpus with itself (1) and can be interpreted as a measure of *self-similarity* of a corpus: the higher the perplexity, the less homogeneous the corpus. Perplexity can be unpacked as the inverse of the corpus probability, normalised by corpus size.

$$PP(W) = P(w_1 \ldots w_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{\prod_{n=1}^{n} P(w_n | w_{1 \ldots n-1})}} \quad (1)$$

We compute the perplexity for language models derived from each of the treebanks. As we are mostly interested in parsing results it is questionable if a simple word trigram model provides the information we are looking for. Hence we computed perplexity[1] for a POS trigram model and for a trigram model based on Leaf-Ancestor (LA) paths (Sampson & Babarczy, 2003). LA measures the similarity of the path of each terminal node in the parse tree to the root node. The path consists

[1] The language models were produced and calculated using the CMU/Cambridge toolkit (`http://mi.eng.cam.ac.uk/~prc14/toolkit.html`)

of the sequence of node labels between the terminal node and the root node, and the similarity of two paths is calculated by using the Levenshtein distance (Levenshtein, 1966). We assume that POS trigrams and LA path representations are more adequate to approximate the syntactic structure of a sentence and to allow predictions about parsing results.

We report experiments on both the randomised and sequential samples. For TüBa-D/Z we have a total of 44 samples with 500 trees each in a 44-cross-validation-style experiment. We compute the perplexity for each of the 44 samples by training a language model on the remaining 43 samples and testing the model on the held-out sample. For TIGER1 and TIGER2 we proceeded as described for TüBa-D/Z.

Table 1 shows that the semantic relatedness in the sequential samples has crucial impact on the size of the vocabulary. We expect that semantic relatedness will lead to a higher predictability of the structure in sequential samples compared to randomised samples, which should result in a lower perplexity for sequential samples. Table 2 shows the results for all samples.

	sequential		*randomised*	
	POS-trigram	**LA-path**	**POS-trigram**	**LA-path**
TIGER1	8.75	6.04	8.88	6.06
TIGER2	8.79	5.89	8.86	6.01
TüBa-D/Z	9.41	4.32	9.43	4.30

Table 2: Perplexity (POS/LA-path-based trigram model) for TIGER and TüBa-D/Z

Results for the POS-trigram and LA-path based models diverge: despite its smaller vocabulary size, the POS-trigram perplexity indicates that TüBa-D/Z is less homogeneous than TIGER, and hence expected to be harder to parse. By contrast, the LA-path based perplexity shows that TIGER (and crucially its annotation scheme as captured by the LA-path based perplexity) is less homogeneous than TüBa-D/Z. In order to resolve this puzzle, in the next section we will investigate the correlation between (POS- and LA-path-based) perplexity and PCFG parsing results.

4 Parsing Experiments

For our parsing experiments we trained the statistical parser BitPar (Schmid, 2004) on our data sets in 44-fold cross-validation-style experiments. For each sample, the training data consists of all remaining samples, so for the first TüBa-D/Z sample we trained the parser on samples 2-44, for sample 2 on samples 1, 3-44 of the treebank, and so forth; and similary for TIGER1 and TIGER2.

4.1 Preprocessing

Before extracting a context free grammar from the treebanks we have to resolve the crossing branches in TIGER. Following Kübler et al. (2006), we resolve the cross-

ing branches by attaching all non-head constituents higher up in the tree. We also include functional labels in the extracted grammars by merging the grammatical function labels with the phrasal nodes or, for terminal nodes, with the POS tags of the node. In TIGER, trees are annotated rather flat in order to capture the semi-free word order of German. This means that while in TüBa-D/Z terminal nodes either have the label HD (head) or the default '-' (non-head), in TIGER terminal nodes comprise a high variety of grammatical functions such as subject, accusative object, dative object, modifier or adpositional case marker. As this would artificially blow up the number of different POS tags in TIGER we inserted unary nodes for all terminal nodes associated with one of the following function labels: SB, OA, DA, AG, OG, OA2 and SBP. The inserted phrasal node carries the grammatical function label of the corresponding terminal node, while the terminal receives the label HD (head). This treebank transformation results in an increased ratio of non-terminal versus terminal nodes to 0.5 (from 0.47) for both TIGER training sets. We then extract a PCFG from each of the training sets and parse our test sets. We evaluate parsing results using `evalb`,[2] an implementation of the PARSEVAL metric, as well as the Leaf-Ancestor (LA) metric (Sampson & Babarczy, 2003).

		TIGER1	TIGER2	TüBa-D/Z
LA (avg.)	*sequential*	88.36	88.45	89.14
	randomised	88.21	88.49	88.95
`evalb`	*sequential*	74.00	73.45	82.80
(≤ 40)	*randomised*	74.33	74.00	83.64

Table 3: Avg. LA and `evalb` results for TIGER and TüBa-D/Z samples

Table 3 shows averaged `evalb` and Leaf-Ancestor (LA) results for the randomised and the sequential samples in our test sets. For all three data sets the `evalb` results for the randomised samples show less variation (min. 71.5 and max. 76.5 for TIGER; min. 80.9 and max. 84.1 for TüBa-D/Z), while the results for the sequential samples are distributed over a wider range from 70 to 79.2 for TIGER and 78 to 85.8 for TüBa-D/Z. `evalb` gives around 10% better results for the parser trained and evaluated on the TüBa-D/Z, while the LA results are much closer across the treebanks within the 88-89% range.

	Perplexity / LA		Perplexity / evalb		sentence length /	
	POS-n-gram	LA-path	POS-n-gram	LA-path	LA	evalb
TIGER1	-0.89	-0.87	-0.76	-0.78	-0.80	-0.78
TIGER2	-0.81	-0.93	-0.81	-0.87	-0.89	-0.81
TüBa-D/Z	-0.47	-0.81	-0.49	-0.74	-0.73	-0.60

Table 4: Pearson's product-moment correlation (sequential samples)

Rehbein & van Genabith (2007) showed that the remarkable difference in `evalb` results for TIGER and TüBa-D/Z is due to the higher ratio of non-terminal vs. terminal nodes in the TüBa-D/Z and that `evalb` cannot be used to compare parsers

[2] All `evalb` results report labelled bracketing f-score.

trained on different treebanks. Therefore we concentrate on the relationship between parsing performance and perplexity (Table 4). For the POS trigram model we compute a strong correlation between perplexity and LA as well as `evalb` parsing results for sequential TIGER samples and a weak correlation for sequential TüBa-D/Z samples. By contrast, the LA-path-based trigram model shows a strong correlation for TIGER and TüBa-D/Z samples. For both models there is no correlation for randomised samples. This means that while for sequential samples a higher perplexity corresponds to lower `evalb` and LA results, this observation does not hold for randomised samples. The same is true for sentence length: while there is a negative correlation between sentence length and parsing results for TIGER samples and, to a lesser extent, for TüBa-D/Z, for randomised samples there is a weak correlation of around -0.45 only. This shows that randomisation succeeded in creating representative samples, where the variation between training and test samples is not high enough to cause differences in parsing results as observed for sequential samples.

4.2 Annotating the TüBa-D/Z in the TIGER Annotation Scheme

In order to conduct a meaningful comparison of the impact of the two different annotation schemes on parser output we extracted a test set of 100 trees from the TüBa-D/Z treebank and manually annotated it following the guidelines in the TIGER stylebook. Due to the high expenditure of time needed for a manual annotation we were able to create a small test set only. To make up for the restricted size we carefully selected our test set by subdividing each of the 44 samples from the TüBa-D/Z treebank into five subsamples with 100 sentences each, and picked the subsample with a sentence length and perplexity closest to the mean sentence length (17.24, mean: 17.27) and mean perplexity computed for the whole treebank (9.44, mean: 9.43). This assures that our test set, despite its limited size, is maximally representative for the treebank as a whole.

We then extracted a training set from the 44 TüBa-D/Z samples (excluding the sentences in the test set). From the TIGER treebank we selected the same number of trees (21898) from the samples 1-44 as well as the first 21898 trees from the samples 45-88 in sequential order and trained the parser on all three training sets (TüBa-D/Z, TIGER1, TIGER2). Then we parsed the test set with the resulting grammars, evaluating the TIGER-trained parser output against the manually created TIGER-style gold-standard of the original TüBa-D/Zstrings and the TüBa-D/Z trained parser output for the same strings against the original TüBa-D/Z trees for those strings. Table 5 shows the parsing results measured with `evalb` and LA.

	TIGER1	TIGER2	TüBa-D/Z
evalb	69.84	71.21	83.35
LA	84.91	86.04	88.94

Table 5: `evalb` and LA results for the manually annotated test set (100 sentences)

As predicted by sentence length and perplexity the LA results for our test set parsed with the TüBa-D/Z grammar is close to the average LA result for the whole TüBa-D/Z (88.95 vs.88.94). For the TIGER grammars parsing TüBa-D/Z-based test strings, however, performance drops from 88.36 to 84.91 (TIGER1) and from 88.45 to 86.04 (TIGER2). The better results for TIGER2 implicate that our TüBa-D/Z-based test set is more similar to the TIGER2 training set, an assumption which is supported by the slightly higher perplexity for TIGER2 compared to TIGER1 (8.79 vs. 8.75), and by the average sentence length for the training sets (TIGER1: 17.96, TIGER2: 17.15, TüBa-D/Z: 17.24). In Section 3.1 we showed that, despite coming from the same domain (newspaper articles, but from two different newspapers), TIGER and TüBa-D/Z are crucially different with regard to the distribution of POS tags, vocabulary size and perplexity. Therefore it is not surprising that the parser trained on a TIGER training set shows lower performance for sentences derived from the TüBa-D/Z.

4.3 Qualitative Evaluation of TIGER and TüBa-D/Z Parser Output

The existence of a small parallel corpus annotated in the TIGER and the TüBa-D/Z annotation schemes allows us to directly compare parser performance for both treebanks. However, the differences in categorial and functional labels used in the annotation often does not support a direct automatic comparison. Hence we focus on the grammatical functions describing the same phenomena in both treebanks. Using the same sentences annotated either in the TIGER or the TüBa-D/Z annotation scheme allows us to assess which functions can be compared. Table 6 gives an overview over some features of our test set a) in the TIGER annotation scheme and b) in the TüBa-D/Z annotation scheme.

	Categorial nodes					*Functional labels*					
	S	NP	AP	PP	AVP	SB	OA	DA	AG	APP	OP
TIGER	155	286	18	164	85	138	67	11	32	12	16
TüBa-D/Z	159	636	105	180	105	140	67	10	0	44	24

Table 6: Overview over some categorial/functional features in both test sets

Table 6 shows that the flat annotation in TIGER leads to a crucially different number of nodes for noun phrases, adjectival phrases and adverbial phrases. The mismatch in the number of PPs is due to the different annotation of pronominal adverbs, which in TüBa-D/Z are always governed by a PP node, while in TIGER only around one-third of the pronominal adverbs projects a PP, while the others are either attached to an S or VP node or, less frequently, to an NP, AP or AVP.

With regard to functional labels there are also considerable differences. While some of the basic argument functions like subjects (SB), accusative objects (OA) and dative objects (DA) follow an approximately similar distribution, most other grammatical functions are interpreted differently in both annotation schemes. One example are appositions (APP): the TüBa-D/Z annotation guidelines consider an

apposition to be an attribute to a noun which has the same case and does not change the meaning of the noun. They do not distinguish between loosely constructed appositions (e.g.: "Angela Merkel, the chancellor") and tightly constructed appositions (e.g.: "the chancellor Angela Merkel") and treat both as appositional constructions. Because of the referential identity of the constituents they do not determine the head of an appositional construction but annotate both constituents as an APP. TIGER only considers loosely constructed appositions which are separated by a comma or another punctuation mark from the preceeding element. Referential identity is also regarded as a constituting property of an apposition, but in contrast to the TüBa-D/Z the first constituent is annotated as the head and the following constituent as an apposition. These differences explain the considerable discrepancy in the number of appositions in both test sets. Another example for the crucial differences in the annotation are pre- and postnominal genitives. In TIGER they are annotated with the label AG, while the same constituents do not get a label in TüBa-D/Z at all and so are not distinguishable from syntactically similar constructions.

	TIGER1			**TIGER2**			**TüBa-D/Z**		
	Prec.	**Recall**	**f-Score**	**Prec.**	**Recall**	**f-Score**	**Prec.**	**Recall**	**f-Score**
subj.	0.64	0.63	0.64	0.66	0.70	0.68	0.73	0.76	0.75
acc. obj.	0.47	0.40	0.43	0.50	0.49	0.50	0.46	0.54	0.50
dat. obj.	0.25	0.18	0.21	0.14	0.09	0.11	0	0	0
conj.	0.47	0.57	0.52	0.44	0.53	0.49	0.53	0.48	0.50
pred.	0.28	0.30	0.29	0.24	0.30	0.27	0.40	0.21	0.28

Table 7: Evaluation of functional labels in the test sets

The functions supporting a direct comparison between both treebanks are subjects, accusative objects, dative, predicates and conjuncts of coordinations (Table 7). The TüBa-D/Z trained parser shows better performance for subjects and comparable results for accusative objects, conjuncts and predicates, while it fails to identify dative objects. But even for grammatical functions which are equally distributed in both treebanks a direct comparison is not straightforward. We will illustrate this for the personal pronoun *es* (it), which often functions as a subject.

The TüBa-D/Z annotation scheme distinguishes three uses of expletive *es*: a) as a formal subject or object without semantic content (eg. for weather verbs), b) as the correlate of an extraposed clausal argument and c) the Vorfeld-*es*. Formal subjects are annotated as subjects, the correlate *es* is either annotated as a subject modifier or a modifier of an object clause, and the Vorfeld-*es*, which is considered to be a purely structural dummy-element, is assigned the label ES. The TIGER annotation scheme also distinguishes three uses of the expletive *es*, but annotates them differently. In TIGER *es* as a formal subject is assigned the label EP instead of the subject label. The Vorfeld-*es* as well as the correlate *es* are both annotated as a placeholder (PH).

This has major consequences for our test sets, where we have 15 personal pronouns with word form *es*. In the TüBa-D/Z annotation scheme 12 of them are

annotated as subjects, the other three as subject modifiers. In TIGER none of them is annotated as a subject. 6 occurrences of *es* are considered to be a placeholder, while the rest is annotated as expletive *es*. If we look at the evaluation results for subjects, 12 of the correctly identified subject relations in the TüBa-D/Z test set are occurrences of expletive *es* (in fact all occurrences of expletive *es* have been assigned the subject label). The linguistic analysis in the TIGER annotation scheme causes more difficulties for the parser to correctly identify the subject. For the placeholders it has to find the corresponding clause and detect the phrase boundaries correctly, which is more challenging than to identify a single token. Another error frequently made by the TIGER grammar is to mistake an expletive *es* as a subject. Here the TüBa-D/Z grammar has a huge advantage as it annotates formal subjects as regular subjects. Caused by the use of an unlexicalised parsing model in some cases the TIGER grammar assigns the label EP to personal pronouns with the word form *er* (he) or *sie* (she). These problems easily explain the gap in evaluation results for subjects between TIGER and TüBa-D/Z and show that even for the same text annotated in the TIGER and in the TüBa-D/Z annotation scheme a fair evaluation is not straightforward at all.

5 Conclusions

We show that due to differences in linguistic analysis as well as out-of-domain problems a direct and fair *automatic* comparison of the TIGER and TüBa-D/Z annotation schemes and their impact on parsing results remains infeasible. There are several attempts to overcome this problem by applying a dependency-based evaluation (Schiehlen, 2004, Versley, 2005), which is considered to be more annotation neutral than labelled bracketing f-scores. A large and detailed dependency-based gold standard for German, the TiGer Dependency Bank (Forst et al., 2004), is also available. Unfortunately the TiGer DB consists of sentences from the TIGER treebank only, thus a fair and unbiased evaluation for text from the TüBa-D/Z can not be guaranteed. Further problems are caused by a mismatch in grammatical functions between the two treebanks and the TiGer DB and the fact, that in TiGer DB auxiliaries are analysed as mere feature carriers and so do not appear in the dependency relations, while in both treebanks they are annotated as the head of the clause. Therefore an extensive conversion of treebank-trained parser output is needed, which is potentially error-prone and it remains unclear to what extent the evaluation results reflect effort in or noise caused by the conversion process. There is only one way out of the dilemma: the creation of a new gold standard which combines test sets from both treebanks, annotated in a dependency-style format independent of the annotation schemes of the original treebanks.

References

[1] Biber, D. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

[2] Brants, S., S. Dipper, S. Hansen, W. Lezius, and G. Smith 2002. The TIGER Treebank. *In Proceedings of the Workshop on Treebanks and Linguistic Theories, Sozopol.*

[3] Corazza, A., A. Lavelli, and G. Satta. Measuring Ambiguity for Parsing Tasks. *Computational Linguistics* (To appear).

[4] Baayen, R. H., H. Van Halteren, and F. Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. In *Literary and Linguistic Computing*, 11: 121-131.

[5] Forst, M., N. Bertomeu, B. Crysmann, F. Fouvry, S. Hansen-Schirra and V. Kordoni. 2004. Towards a dependency-based gold standard for German parsers - The TiGer Dependency Bank. In *Proceedings of the LINC-04 Workshop*, Geneva, Suisse.

[6] Höhle, T. 1998. Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. *In Akten des Siebten Internationalen Germansitenkongresses 1985*, 329-340, Göttingen, Germany.

[7] Kilgariff, A. Comparing Corpora. *International Journal of Corpus Linguistics* 6(1): 1-37.

[8] Kübler, S., E. Hinrichs, and W. Maier. 2006. Is it Really that Difficult to Parse German? *In Proceedings of the EMNLP-CoNLL 2006)*, Sydney, Australia.

[9] Levensthein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady*, 10.707-10.

[10] Nerbonne, J. and W. Wiersma 2006. A Measure of Aggregate Syntactic Distance. *In Proceedings of the Workshop on Linguistic Distances*, CoNLL & ACL, Sydney, Australia.

[11] Rehbein, I. and J. v. Genabith. 2007. Treebank Annotation Schemes and Parser Evaluation for German. *In Proc. of the EMNLP-CoNLL 2007*, Prague, Czech Republic.

[12] Sampson, G., and A. Babarczy. 2003. A test of the leaf-ancestor metric for parse accuracy. *Natural Language Engineering*, 9 (4): 365-380.

[13] Sanders, N. C. 2007. Measuring Syntactic Differences in British English. *In Proceedings of the COLING/ACL 2007 Student Research Workshop)*, Prague, Czech Republic.

[14] Schiehlen, M. 2004. Annotation Strategies for Probabilistic Parsing in German. *In Proceedings of the COLING 2004*, Geneva, Switzerland.

[15] Schmid, H.. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. *In Proceedings of the COLING 2004*, Geneva, Switzerland.

[16] Telljohann, H., E. W. Hinrichs, S. Kübler, and H. Zinsmeister. 2005. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).* Universität Tübingen, Germany.

[17] Versley, Y. 2005. Parser Evaluation Across Text Types. In *Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT 2005)*, Barcelona, Spain.

[18] Schiller, A., S. Teufel, and C. Thielen. 1995. *Guidelines für das Tagging deutscher Textcorpora mit STTS.* Technical Report, IMS-CL, University Stuttgart.

[19] Ziegler, A., K. Best, and G. Altmann. 2002. Nominalstil. In: *Empirische Text- und Kulturforschung* 72-85. RAM-Verlag (RAM-Verlag@t-online.de)